Fiche TD avec le logiciel  : ter1

Représentations triangulaires

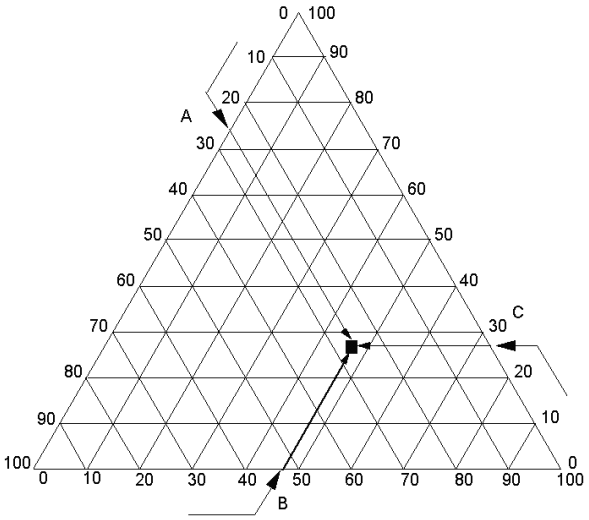
D. Chessel, A.B. Dufour & J.R. Lobry

La représentation triangulaire place dans un triangle équilatéral un point pour représenter une distribution de fréquences sur trois catégories. Cette pratique élémentaire permet d'illustrer sans artifice quelques idées fondamentales de la statistique euclidienne.

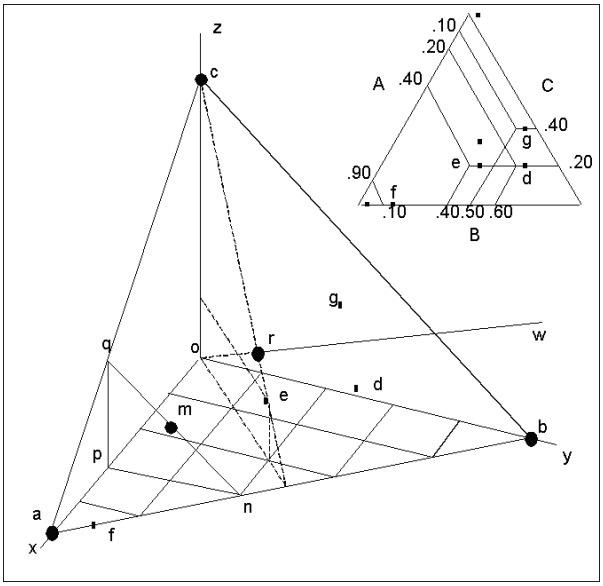
Table des matières

1 Définitions

La représentation triangulaire est le procédé graphique par lequel on place un point à trois coordonnées (a, b, c) positives ou nulles et vérifiant $a+b+c = 1$ ou $a + b + c = 100$ dans un triangle équilatéral, conformément au principe :



Ce procédé déguise sous une apparence pragmatique le fait que le point (x, y, z) appartient au plan $x+y+z = 1$ à l'intérieur du triangle défini par les extrémités des vecteurs de la base canonique $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$ et $\mathbf{e}_3 = (0, 0, 1)$ représenté par les trois points a, b et c :



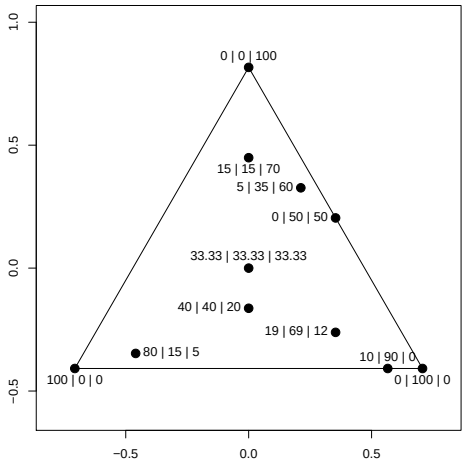
Le premier exercice consiste à écrire une fonction qui dessine le triangle et place dedans quelques points. Il suffit de trouver une base orthonormée de $\mathbb{R}^3$ dont le premier vecteur est perpendiculaire au plan. Il ne s'agit ici que de géométrie ordinaire.

Le vecteur $\mathbf{u} = (1,1,1)/\sqrt{3}$ convient. On prendra le second perpendiculaire au premier dans le plan $z = 0$. Il y a deux solutions et on prend $\mathbf{v} = (-1/\sqrt{2}, 1/\sqrt{2}, 0)$. Le troisième doit être normé et perpendiculaire au précédent. $\mathbf{w} = (-1/\sqrt{6}, -1/\sqrt{6}, 2/\sqrt{6})$ convient. Les coordonnées d'un point quelconque dans la base canonique sont (x, y, z). Dans la nouvelle base les coordonnées du même point sont $(0, (y-x)/\sqrt{2}, (2z-x-y)/\sqrt{6})$.

L'essentiel est fait. Pour vous aider :

```
library(ade4)
foo1 <- function() {
    a <- c(-1/sqrt(2), -1/sqrt(6))
    b <- c(1/sqrt(2), -1/sqrt(6))
    c <- c(0, 2/sqrt(6))
    w <- rbind(a, b, c)
    w
}
foo2 <- function() {
    w <- foo1()
    plot(w, asp = 1, xlim = c(-0.8, 0.8), xlab = "", ylab = "")
    polygon(w)
}
foo3 <- function(f, pos = 4) {
    f <- f/sum(f)
    w <- foo1()
    wf <- apply(w * f, 2, sum)
    points(wf[1], wf[2], pch = 19, cex = 1.5)
    f <- round(100 * f, dig = 2)
    text(wf[1], wf[2], paste(f[1], f[2], f[3], sep = " | "), pos = pos)
}
```

Refaites cette figure :



Ce qui est essentiel tient dans la figure 1 : le triangle et la position du point dans le triangle ont plusieurs interprétations.

- ⋆ La première (en haut, à gauche) est traditionnelle. Des segments parallèles aux côtés divisent chacun d'entre eux dans l'échelle 0-100 et chaque point $x + y + z = 100$ prend sa place par trois droites concourantes.
- ⋆ La seconde (en haut, à droite) est numérique. Deux coordonnées place un point. Tout vient du fait que trois coordonnées sont redondantes, puisque deux d'entre elles dans $x + y + z = 1$ impose une valeur à la troisième. Cette réduction de dimension, ici sans perte d'information, se fera ensuite avec une perte obligatoire mais minimale. Les coordonnées sont des scores.

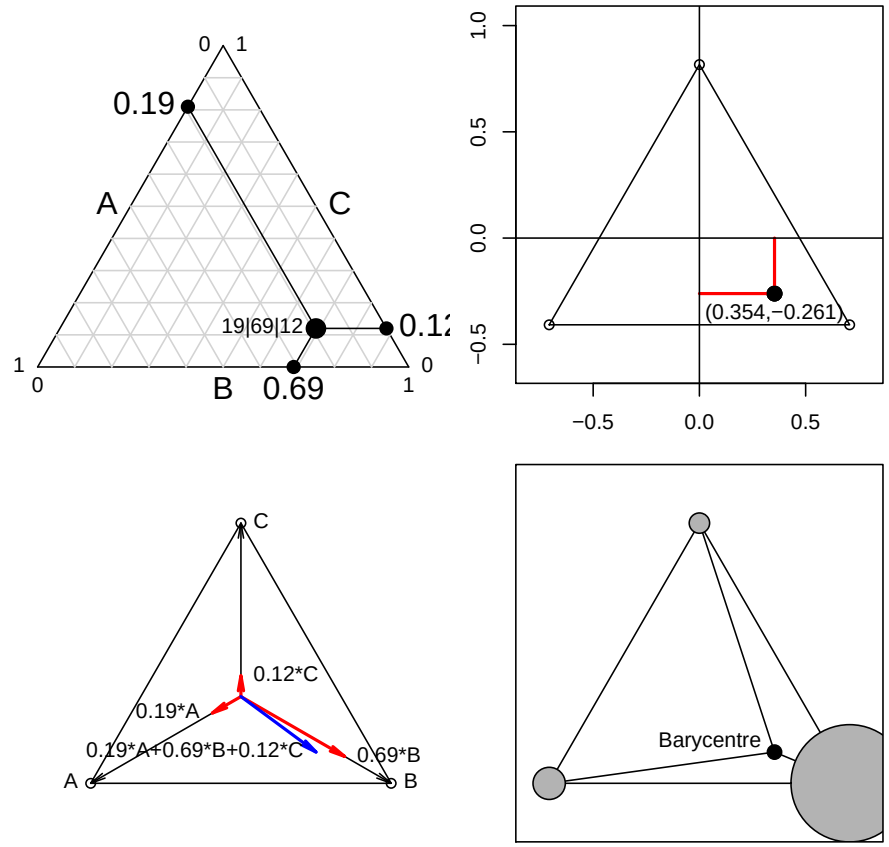


FIG. 1 – Quatre principes de lecture d'une représentation triangulaire

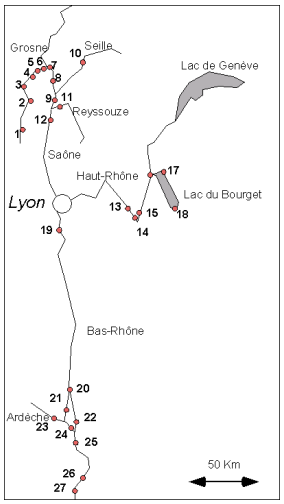
⋆ La troisième, en bas à gauche est algébrique et traduit $(x, y, z) = x(1, 0, 0) + y(0, 1, 0) + z(0, 0, 1)$. Rien n'empêche d'utiliser une rotation, une symétrie, un triangle non équilatéral, pour garder un sens à cette opération. Ces choix seront autant d'intentions pour orienter le lecteur.

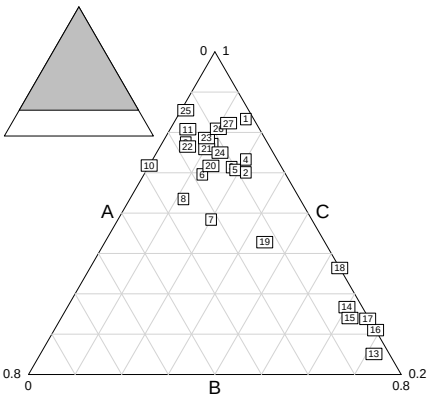
⋆ La dernière est mécanique. Le point représenté est le centre de gravité de trois points affectés des masses x, y et z. Les catégories et les mesures, les individus et les variables se réunissent pour former des *biplots*. Aucune interprétation n'a plus de valeur qu'une autre. Ce sont des principes qui définissent une figure et qui définiront, dès la dimension 4, de multiples extensions possibles.

2 Deux dimensions contre deux dimensions

Quand on cherche le lien entre la variable `x` et la variable `y`, la première chose à faire est `plot(x,y)`. Mais une feuille de papier a deux dimensions. Quand l'information est formée d'un couple d'informations comportant plusieurs dimensions, c'est plus compliqué. Supposons qu'on ait deux dimensions dans chaque élément. Les données sont dans [1]. On dispose de 27 stations réparties le long du bassin rhôdanien conformément à la carte ci-dessous, à gauche. Dans chacune des stations des poissons (chevaines, *Leuciscus cephalus*) sont typés pour 4 enzymes polymorphes. Seul le premier gène comporte trois allèles présents et nous intéresse ici. Cela donne la représentation triangulaire, à droite. Le même jeu d'étiquettes permet de se rendre compte immédiatement qu'il existe un lien fort entre les deux types d'information : ce lien, expression du fonctionnement des populations à travers les échanges géniques et les structures en résultant, est au centre de l'intention expérimentale.

```
data(chevaine)
t3locus <- chevaine$tab[, 1:3]
names(t3locus) <- c("A", "B", "C")
triangle.plot(t3locus, clab = 0.75)
```



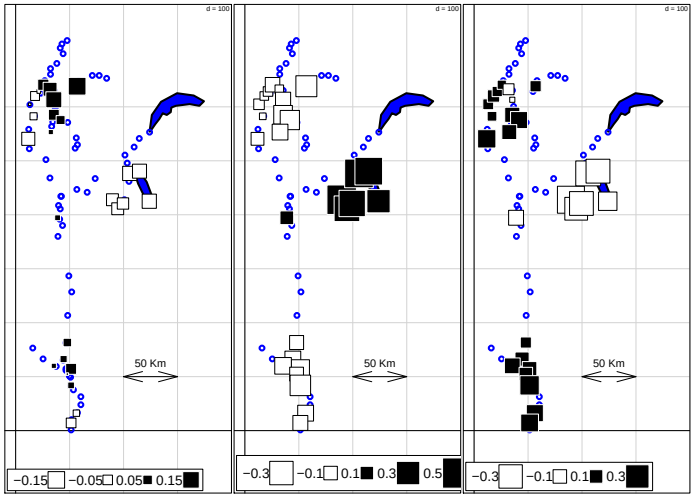


On devra alors choisir entre représenter les données dans l'espace concret ou l'espace concret dans les données. On obtient des objets techniquement très différents devant conduire –on l'espère– à une interprétation commune. Il suffit d'examiner la représentation triangulaire pour voir que l'essentiel de la variabilité est associée à l'originalité du groupe des stations du Haut-Rhône. On dit que le Rhône est biologiquement un affluent de la Saône.

La même constatation s'impose à la lecture des trois cartes représentant chacune des fréquences alléliques. Chacune des figures a ses propriétés propres et ne saurait systématiquement s'imposer à l'autre. Chercher à refaire la figure ci-dessous à l'aide de la documentation de l'objet `chevaine` en prenant soin

1. de centrer les fréquences (`?scale`), d'une part ;
2. de garantir à chaque fenêtre la même échelle, d'autre part (voir `s.value`).

Expliquer en quoi ce deuxième point est important et indiquer en quoi il permet de mieux appréhender la représentation triangulaire.

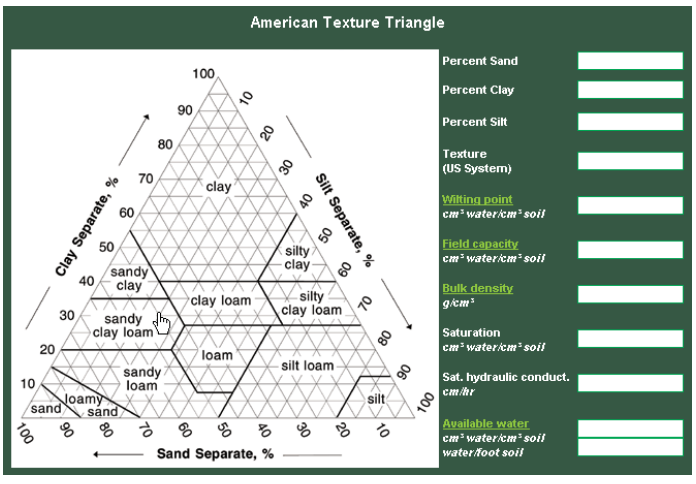


3 Argile, limon, sable

La représentation triangulaire est bien connu des agronomes. Sur `http://www.geo.unizh.ch/virtualcampus/doityoursoil/f/module_2/sequence_10/infos_2_10_f.html` :

> La texture est une propriété du sol qui traduit de manière globale la composition granulométrique de la terre fine. Elle reflète la part respective des constituants triés selon leur taille et est déterminée dans un *triangle des textures*. On distingue la texture minérale qui est la proportion des sables, limons et argiles, et la texture organique qui reflète la proportion de fibres et de matériel fin microagrégé dans les matériaux holorganiques (tourbes, litières, composts).

Le triangle des textures a une version américaine :



`http://www.pedosphere.com/resources/texture/triangle_us.cfm`

Quand on clique sur le triangle, le site redonne une distribution de fréquence. On peut écrire une fonction qui fait cela après un `triangle.plot` :

```
w <- matrix(c(1, 0, 0, 0, 1, 0, 0, 0, 1), 3)
w = as.data.frame(w)
triangle.plot(w)
f <- function(loc) {
    a <- loc$x
    b <- loc$y
    z <- 1/3 + sqrt(6) * b/3
    x <- z - b * sqrt(6)/2 - a/sqrt(2)
    y <- 1 - x - z
    return(round(c(x, y, z), dig = 2))
}
f(locator(1))
```

Remarquer que les flèches tournent dans le sens des aiguilles d'une montre et que les variables peuvent permuter. C'est déjà l'indication qu'une carte factorielle n'a pas d'orientation canonique. La version canadienne n'a conservé que deux des trois variables (`http://www.pedosphere.com/resources/bulkdensity/triangle.cfm`).

4 Rouge, vert, bleu

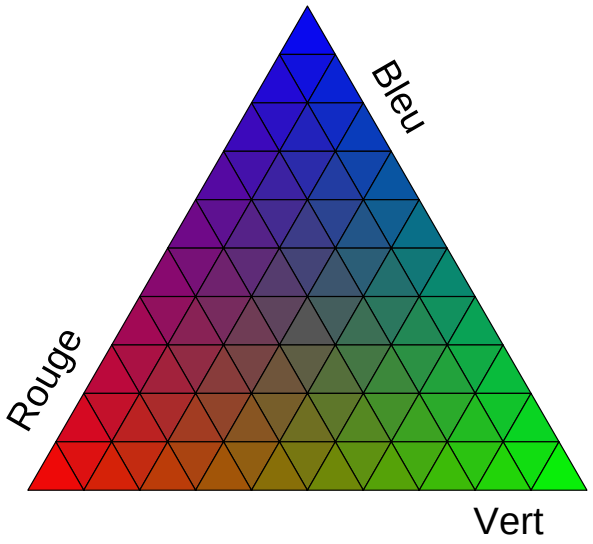
Dans la synthèse additive des couleurs (`http://www.irht.cnrs.fr/formation/cours/acq/visualisation.htm`), la base est formée du triplet Rouge-Vert-Bleu (*RGB*) et un des codages numériques fondamentaux de la couleur est l'écriture d'une proportion des trois fondamentaux :

Laurence Jacquet `www.redisdead.net/rvb/RVB.png`

La fonction `rgb()` permet de créer une couleur comme une distribution de fréquences en Rouge-Vert-Bleu.

```
tricolor <- function() {
    tri <- function(x) {
        x <- x/sum(x)
        return(c((x[2] - x[1])/sqrt(2), (2 * x[3] - x[2] - x[1])/sqrt(6)))
    }
    opar <- par(mar = par("mar"))
    on.exit(par(opar))
    A <- tri(c(1, 0, 0))
    B <- tri(c(0, 1, 0))
    C <- tri(c(0, 0, 1))
    par(mar = c(0.1, 0.1, 0.1, 0.1))
    plot(0, 0, type = "n", xlim = c(-0.8, 0.8), ylim = c(-0.6, 1),
        xlab = "", ylab = "", xaxt = "n", yaxt = "n", asp = 1, frame.plot = FALSE)
    text(A[1] + 0.15, A[2] + 0.4, labels = "Rouge", cex = 2, pos = 2,
        srt = 60)
    text(B[1] - 0.2, B[2] - 0.025, labels = "Vert", cex = 2, pos = 1)
    text(C[1] + 0.15, C[2] - 0.15, labels = "Bleu", cex = 2, pos = 4,
        srt = -60)
    for (i in 1:10) {
        a <- (i - 1)/10
        b <- 0
        w <- c(a, b)
        dir = TRUE
        l0 <- c(NA, NA)
        l1 <- c(NA, NA, NA)
        while ((0 <= sum(w)) & (sum(w) <= 1)) {
            l1 <- rbind(l1, c(w[1], w[2], 1 - sum(w)))
            lll <- nrow(l1)
            y <- tri(l1[lll, ])
            l0 <- rbind(l0, y)
```

```
            if (dir) {
                w <- w + c(0.1, 0)
            }
            else {
                w <- w + c(-0.1, 0.1)
            }
            dir <- !dir
            if (lll > 3) {
                colo <- apply(l1[lll:(lll - 2), ], 2, mean)
                colo <- rgb(colo[1], colo[2], colo[3])
                polygon(x = l0[lll:(lll - 2), 1], y = l0[lll:(lll -
                  2), 2], col = colo)
            }
        }
    }
}
tricolor()
```

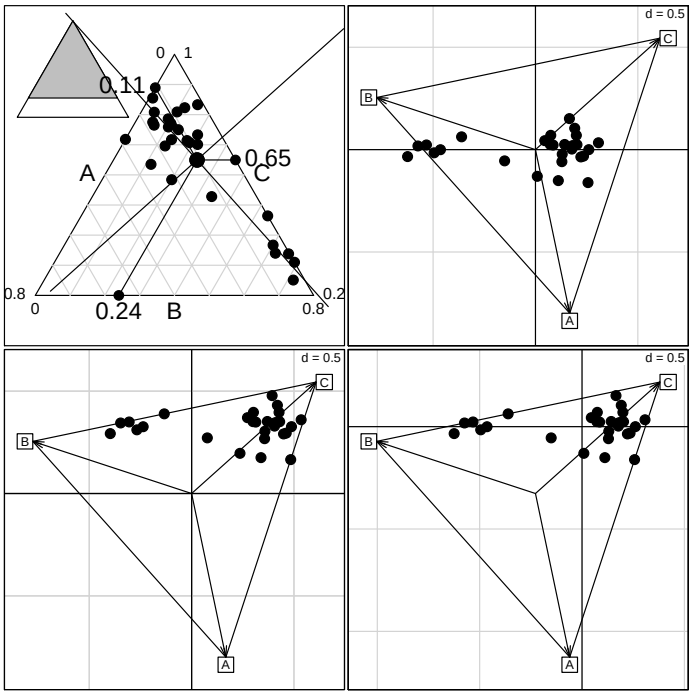


5 La position de l'origine

L'intérêt pédagogique de la représentation triangulaire est considérable. Elle permet de poser les problèmes de l'analyse des données à deux dimensions – donc de manière visible– sans pour autant dégrader ces questions en trivialité. C'est en effet un des paradoxes de cette discipline que de devoir expliciter ses procédures sur des exemples accessibles et de s'en servir effectivement sur des jeux de données qui ne le sont pas. En d'autres termes, on a le choix de montrer soit comment ça marche, soit à quoi ça sert, mais on peut difficilement faire les deux à la fois. Parmi les questions élémentaires figure celle de la position de l'origine. L'analyse des données vise à la représentation des typologies, donc des modes d'organisation des objets. Ce qui compte dans une représentation triangulaire, c'est la position relative d'un point par rapport aux trois sommets, puis la position relative d'une point par rapport à chacun des autres. Il n'existe pas à

proprement parler une origine naturelle dans la figure, un point de coordonnées $(0, 0)$ mais une obligation technique de placer une origine pour tracer le dessin. Examiner les quatre éléments de la figure suivante.

1. Le premier –en haut, à gauche– redonne la représentation triangulaire du paragraphe 2 à laquelle on a rajouté le centre de gravité et les axes principaux du nuage de points. Si ces termes semblent mal définis, on peut se contenter pour le moment de voir la fonction de ces éléments : placer une position centrale du nuage et indiquer quelle en est la direction privilégiée de variabilité. Aucune référence à une quelconque origine ne figure sur le graphe, mais par convention l'origine est au centre du triangle.
2. Le second –en haut, à droite– représente le même nuage de points. Mais par convention l'origine est placée au centre de gravité (point moyen) du nuage. Ce déplacement qui correspond au changement $X = x - \bar{x}$ $Y = y - \bar{y}$ et $Z = z - \bar{z}$ ne change strictement rien aux positions relatives (translation). On utilise les axes principaux comme nouveau repère, ce qui correspond à une simple rotation et on ajoute une représentation de la base canonique. On a perdu la relation d'*averaging* et les points ne sont plus à la moyenne de la distribution de poids qu'ils définissent.
3. On veut alors restaurer cette propriété. Si on garde la position des catégories, définie par la base canonique, le nuage de points se déplace par une translation inverse. La direction principale est conservée (en bas à gauche).
4. On opère enfin une translation du système de référence pour rendre au *biplot* sa fonction de représentation triangulaire. Le graphe ainsi construit est strictement une représentation par *averaging* quand il y a trois allèles mais s'étendra à un nombre quelconque de variables en conservant cette propriété.



Pour refaire cette figure :

```
op <- par(no.readonly = TRUE)
par(mfrow = c(2, 2))
par(mar = c(0.1, 0.1, 0.1, 0.1))
triangle.plot(t3locus, addm = T, adda = T, cpoi = 2.5, box = T)
w = dudi.pca(t3locus, scal = F, scan = F)
s.chull(w$c1, factor(rep(1, 3)), clab = 0)
s.arrow(w$c1, add.p = T)
s.label(w$li, clab = 0, cpoi = 2.5, add.p = T)
s.arrow(w$c1)
s.chull(w$c1, factor(rep(1, 3)), add.p = T, clab = 0)
points(as.matrix(t3locus) %*% as.matrix(w$c1), pch = 20, cex = par("cex") *
    2.5)
s.multinom(w$c1, t3locus, n.sample = 0, clabelrowprof = 0, cpointrowprof = 2.5,
    pchrowprof = 20, translate = TRUE)
par(op)
```

6 Bas, moyen, haut

On connaît les variables quantitatives (ou numériques) qui mesurent une quantité à l'aide d'une unité, les variables qualitatives ou facteurs qui prennent leur valeur dans un ensemble de modalités, les variables ordinales ou facteur à modalités ordonnées qui forment une transition entre les deux précédentes. La représentation triangulaire introduit aux variables distributionnelles : la distribution de fréquences peut être une distribution de poids, d'effectif, de surface, de valeur, etc. Utiliser trois dimensions c'est déjà enregistrer une variable sans totalement préjuger de la forme de variation de cette variable. Par exemple, dans [3], le relief d'un district naturel est enregistré sous forme du pourcentage de la surface en trois catégories :

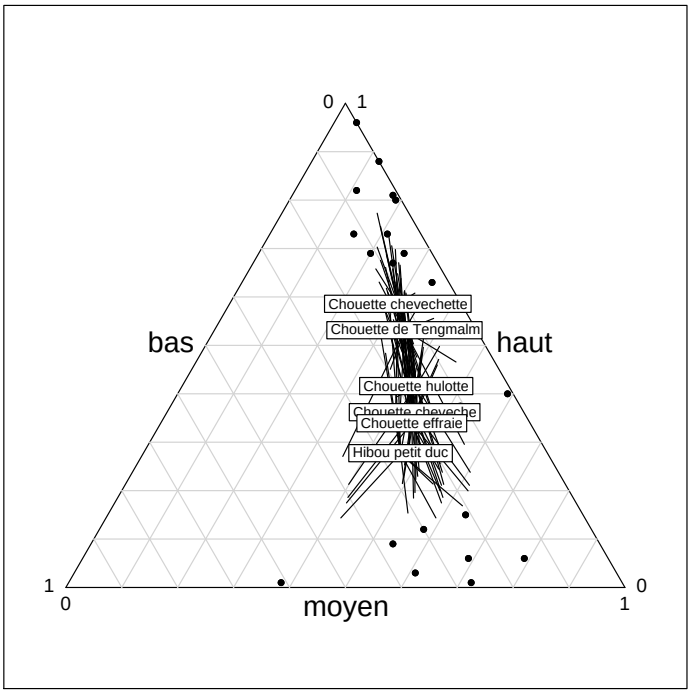
bas plaines et collines],800 m]

moyen moyenne-montagne]800-1500 m]

haut haute-montagne]1500m,]

On considère la distribution de six espèces de rapaces nocturnes :

```
data(atlas)
w1 <- as.data.frame(triangle.plot(atlas$alti, box = F, show = F))
w2 <- atlas$birds[, 9:14]
lab <- gsub("_", " ", names(atlas$birds)[9:14])
s.distri(w1, w2, add.p = T, cell = 0, axesell = F, cstar = 0.5,
    clab = 0.75, lab = lab)
```



On remarquera la question des dimensions. Les données d'environnement (pourcentages) sont à trois composantes. Mathématiquement, la dimension est deux, la somme des trois variables est constante. Expérimentalement la dimension est voisine de un, les points étant presque tous sur une droite. Les espèces se répartissent sur le gradient d'altitude. Repérer et séparer les contraintes mathématiques et les particularités expérimentales est le rôle de l'analyse des données. Ici la variable distributionnelle cache une variable quantitative simple. Ce n'est pas toujours le cas, évidemment.

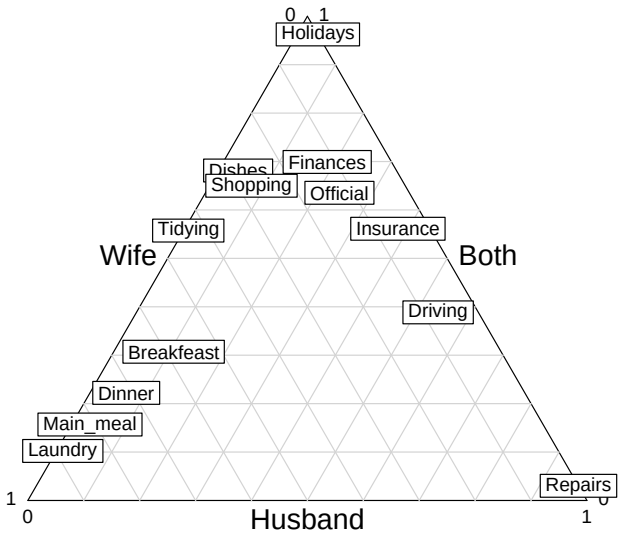
7 Elle, lui, les deux

Un exemple, qui pourrait être sans commentaire : une enquête auprès des ménages rapportée dans [2] demande à un couple : 'Qui assure chacune des tâches ci-dessous ?' La réponse peut être Monsieur, Madame, Ensemble ou Alternativement. Les données contiennent les pourcentages de réponse.

```
data(housetasks)
hc <- t(apply(housetasks, 1, function(x) x/sum(x)))
hc <- cbind.data.frame(Wife = hc[, 1], Husband = hc[, 3], Both = hc[,
    2] + hc[, 4])
t(round(hc, dig = 2))
```

```
        Laundry Main_meal Dinner Breakfeast Tidying Dishes Shopping Official Driving
Wife       0.89      0.81   0.71       0.59    0.43   0.28     0.28     0.12    0.07
Husband    0.01      0.03   0.06       0.11    0.01   0.04     0.08     0.24    0.54
Both       0.10      0.16   0.22       0.31    0.56   0.68     0.65     0.64    0.39
        Finances Insurance Repairs Holidays
Wife        0.12      0.06    0.00     0.00
Husband     0.19      0.38    0.97     0.04
Both        0.70      0.56    0.03     0.96
```

```
triangle.plot(hc, show = F, clab = 1, label = row.names(hc))
```



8 La variabilité multinomiale

Parmi les variables distributionnelles, les distributions mesurées en effectifs sont particulières. La distribution sur trois catégories peut être le résultat d'un tirage multinomial. C'est le cas, en génétique, sous l'hypothèse de Hardy-Weinberg. Une population présente, pour un locus polymorphe, des fréquences alléliques du type (p_1, p_2, p_3). Sélectionner m individus, revient à tirer au hasard $2m$ allèles et la distribution des effectifs observés (X_1, X_2, X_3) suit une loi multinomiale de paramètres $2m$ et (p_1, p_2, p_3). La position du point représentatif de l'échantillon sur le triangle suit alors, approximativement, une loi normale.

Considérons quatre profils théoriques, respectivement $(0.49, 0.47, 0.04)$, $(0.4, 0.4, 0.3)$, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ et $(0.05, 0.70, 0.25)$. Pour refaire le premier élément de la figure 2 :

```
par(mar = c(0.1, 0.1, 0.1, 0.1))
proba <- matrix(c(0.49, 0.47, 0.04, 0.4, 0.3, 0.3, 0.05, 0.05, 0.9,
    0.05, 0.7, 0.25), ncol = 3, byrow = T)
proba.df <- as.data.frame(proba)
names(proba.df) <- c("A", "B", "C")
row.names(proba.df) <- c("P1", "P2", "P3", "P4")
w.proba <- triangle.plot(proba.df, clab = 2, show = F)
box()
```

Le triangle contient la position théorique du profil. Si on choisit maintenant un certain nombre d'échantillons aléatoires de ce profil, on peut représenter la position observée. Par exemple prenons des tailles d'échantillons respectives de 50, 30, 20 et 80 individus. Simulons 100 fois un tirage aléatoire, calculons les fréquences observées et la position du point dans le triangle. Les pétales de *sunflower* indiquent les points superposés.

```
triangle.plot(proba, clab = 0, show = F, cpoi = 0, min3 = rep(0,
    3), max3 = rep(1, 3))
w.tri = data.frame(x = c(-sqrt(1/2), sqrt(1/2), 0), y = c(-1/sqrt(6),
    -1/sqrt(6), 2/sqrt(6)))
L3 <- c("A", "B", "C")
row.names(w.tri) <- L3
w <- as.matrix(w.tri)
n1 <- function(nrepet, nsample, proba, color) {
    tab <- apply(rmultinom(nrepet, nsample, proba), 2, function(x) x/sum(x))
    xypro <- t(tab) %*% w
    sunflowerplot(xypro, add = T, seg.col = "black", rot = TRUE)
    points(xypro[, 1], xypro[, 2], pch = 19, col = color, cex = 1)
    return(xypro)
}
xypro1 <- n1(100, 50, proba[1, ], "indianred1")
xypro2 <- n1(100, 30, proba[2, ], "yellow3")
xypro3 <- n1(100, 80, proba[3, ], "lawngreen")
xypro4 <- n1(100, 20, proba[4, ], "steelblue3")
par(mar = c(0.1, 0.1, 0.1, 0.1))
box()
```

Chacun des nuages de réalisations donne une matrice de variances-covariances et une ellipse d'inertie qui résume la dispersion des simulations.

```
triangle.plot(proba, clab = 0, show = F, cpoi = 0, min3 = rep(0,
    3), max3 = rep(1, 3))
xypro <- rbind(xypro1, xypro2, xypro3, xypro4)
xyfac <- as.factor(rep(1:4, rep(100, 4)))
s.class(xypro, xyfac, add.plot = T, col = c("indianred1", "yellow3",
    "lawngreen", "steelblue3"), cell = 2.45)
```

Supposons enfin que nous n'ayons qu'un échantillon observé.

```
nsample <- c(50, 30, 80, 20)
obs.df <- as.data.frame(matrix(c(21, 15, 5, 3, 25, 7, 4, 10, 4,
    8, 71, 7), nrow = 4))
row.names(obs.df) <- c("O1", "O2", "O3", "O4")
names(obs.df) <- c("A", "B", "C")
w.mul <- s.multinom(w.tri, obs.df, coul = c("indianred1", "yellow3",
    "lawngreen", "steelblue3"), translate = F)
points(sweep(w.proba, 2, w.mul$tra), cex = 2, pch = 20)
```

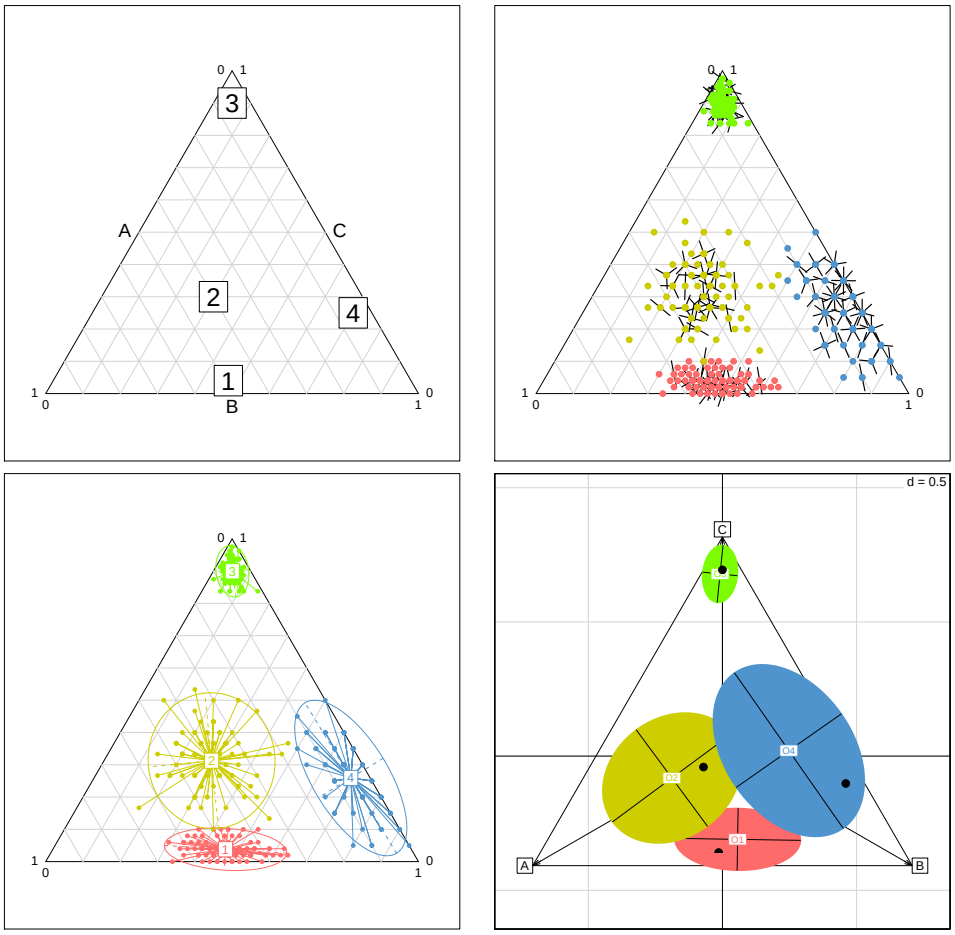


FIG. 2 – Représentation triangulaire et variabilité multinomiale. En haut, à gauche : position dans le triangle de quatre distributions de fréquences à trois composantes. A côté : positions dans le triangle de 100 échantillons aléatoires simples de chacune des distributions. Les profils 1, 2, 3 et 4 sont représentés respectivement par des échantillons de taille respective 50, 30, 80 et 20 individus. Les pétales représentent les superpositions (fonction *sunflowerplot*). En dessous, Les mêmes échantillons donnent une ellipse de dispersion qui contient approximativement 95% des points d'un groupe. La quatrième fenêtre donne, à l'inverse la position attendue du vrai profil à partir d'un échantillon. Dans 95 % des cas, cette position est à l'intérieur de l'ellipse (zone de confiance). Faire des essais variés et observer que cette pratique est invalidée par l'absence d'une catégorie dans l'échantillon. La fonction *s.multinom* place l'origine au centre de gravité de l'ensemble des échantillons

9 Primaire, secondaire, tertiaire

La représentation triangulaire a une importance pédagogique considérable. Elle permet de poser les questions qu'aborde l'analyse des données dans un cadre accessible mais non trivial. L'analyse multivariée ne commence pas avec deux variables. Entre une régression simple et une régression multiple, il y a un décalage technique (la première est le cas particulier de la seconde) mais une identité de logique et d'objectif.

L'analyse multivariée commence quand on substitue la question de la typologie (ce qui se ressemble et ce qui est différent) à la question de la valeur (observée et prédite). Dépouiller une analyse typologique, c'est faire une synthèse des différences deux à deux dans un ensemble d'objets et en préciser l'origine. De ce point de vue, dans les données ternaires (profils à trois catégories) la représentation triangulaire assure complètement et définitivement cette tâche. Il suffit de lire la figure pour parler des données.

Ce qui vient ici s'ajouter c'est que la typologie porte sur des objets dont une partie des différences deux à deux a du sens et qu'une autre partie n'en n'a pas. On devra donc orienter la lecture.

```
data(euro123)
w <- cbind.data.frame(cbind.data.frame(euro123$in78, euro123$in86,
    euro123$in97))
names(w) <- paste(c("pri", "sec", "ter"), rep(c("78", "86", "97"),
    c(3, 3, 3)), sep = "")
```

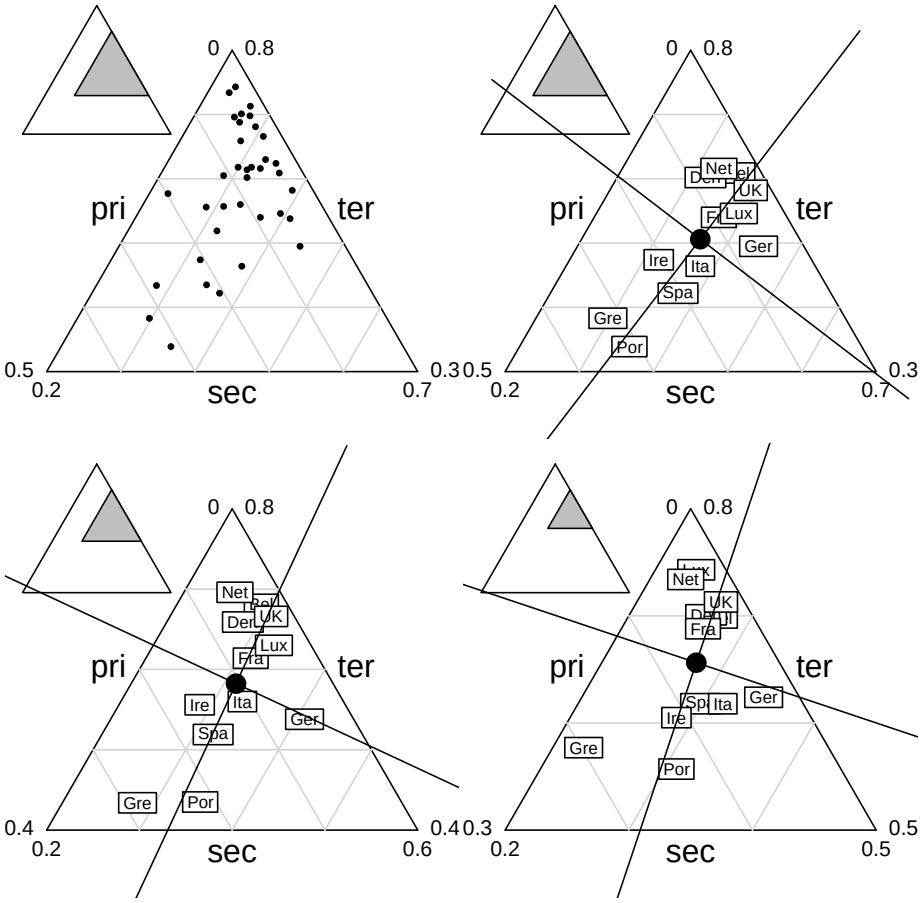
On a pour 12 pays de la communauté économique européenne la distribution des emplois dans les catégories classiques (primaire, secondaire et tertiaire), ceci pour les années 1978, 1986 et 1997.

	1978			1986			1997		
	pri	sec	ter	pri	sec	ter	pri	sec	ter
Belgium	0.03	0.36	0.61	0.03	0.29	0.68	0.03	0.28	0.70
Denmark	0.08	0.32	0.60	0.06	0.28	0.66	0.04	0.26	0.70
Spain	0.21	0.37	0.42	0.16	0.32	0.52	0.08	0.30	0.62
France	0.09	0.37	0.54	0.07	0.31	0.61	0.05	0.27	0.69
Greece	0.32	0.30	0.38	0.28	0.28	0.43	0.20	0.23	0.58
Ireland	0.21	0.32	0.47	0.16	0.29	0.56	0.11	0.29	0.60
Italy	0.16	0.38	0.46	0.11	0.33	0.56	0.07	0.32	0.62
Luxembourg	0.06	0.39	0.55	0.04	0.33	0.63	0.02	0.23	0.74
Netherlands	0.05	0.33	0.62	0.05	0.26	0.70	0.04	0.23	0.73
Portugal	0.31	0.35	0.34	0.22	0.35	0.44	0.13	0.31	0.56
Germany	0.06	0.44	0.50	0.05	0.41	0.54	0.03	0.35	0.62
United Kingdom	0.03	0.39	0.58	0.03	0.31	0.67	0.02	0.27	0.71

Comparer deux secteurs économiques en un pays à une date donnée a un sens ; comparer un secteur économique dans deux pays à une date aussi, comparer un secteur dans un pays entre deux dates également. On voit immédiatement que dans la variabilité totale une part est associée aux secteurs d'activité, une part au temps et une part à l'histoire de chaque nation.

On peut se centrer sur la stabilité de la typologie entre pays :

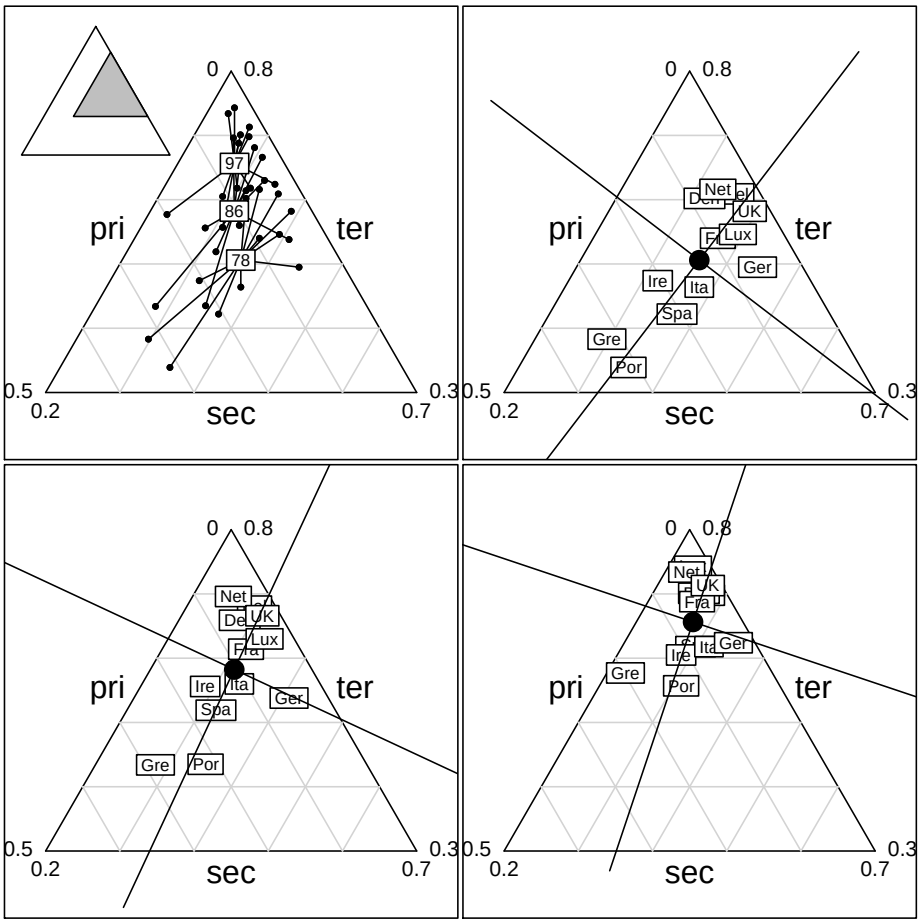
```
par(mfrow = c(2, 2))
x = rbind.data.frame(euro123$in78, euro123$in86, euro123$in97)
w = triangle.plot(x)
triangle.plot(euro123$in78, clab = 1, label = unique(euro123$plan$pays),
    adda = T)
triangle.plot(euro123$in86, clab = 1, label = unique(euro123$plan$pays),
    adda = T)
triangle.plot(euro123$in97, clab = 1, label = unique(euro123$plan$pays),
    adda = T)
```



Regarder, comme le Luxembourg prend progressivement la tête dans la course à la tertiarisation alors que l'Allemagne garde son statut de pays industrialisé, comment les groupes de pays restent assez stable alors que l'ensemble bouge. Les questions d'échelles sont renvoyées dans le cartouche (en haut, à gauche) et l'image se concentre sur la typologie intra dates et inter-pays. Mais on peut ramener les figures dans un cadre unique :

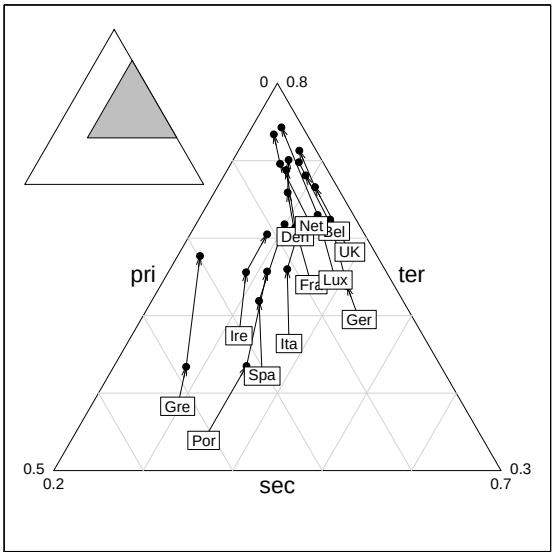
```
par(mfrow = c(2, 2))
w = triangle.plot(x)
s.class(w, euro123$plan$an, cell = 0, add.p = T)
lab0 <- unique(euro123$plan$pays)
m0 <- c(0, 0.2, 0.3)
m1 <- c(0.5, 0.7, 0.8)
triangle.plot(euro123$in78, clab = 1, label = lab0, min3 = m0, max3 = m1,
    adda = T, box = T, show = F)
```

```
triangle.plot(euro123$in86, clab = 1, label = lab0, min3 = m0, max3 = m1,
    adda = T, box = T, show = F)
triangle.plot(euro123$in97, clab = 1, label = lab0, min3 = m0, max3 = m1,
    adda = T, box = T, show = F)
```



On voit maintenant le temps en action dans le déplacement global du nuage de points et dans la réduction progressive de la variabilité. La diversité des questions de la statistique multi-tableau est explicite : ce qui reste stable ou évolue peut concerner chaque point, le point moyen, l'hétérogénéité du nuage de points, les positions relatives entre points, les axes principaux, la typologie.

```
par(mfrow = c(1, 1))
w = triangle.plot(x, cpo = 1.5)
s.match(w[euro123$plan$an == "78", ], w[euro123$plan$an == "86",
    ], add.p = T, clab = 0)
s.match(w[euro123$plan$an == "86", ], w[euro123$plan$an == "97",
    ], add.p = T, clab = 0)
s.label(w[euro123$plan$an == "78", ], add.p = T, lab = lab0)
```



C'est bien toujours la même figure, mais pas la même lecture. Dès que le nombre de dimensions sera plus grand que 2, il faudra en plus choisir des méthodes différentes. Seule la dimension 2 permet ce genre d'exercice.

10 Exercice : Régulièrement, occasionnellement ou jamais ?

Les données triangulaires sont utilisées dans les enquêtes publiques. L'exercice proposé ici est issu d'un rapport de Corinne Jehl et Camille Mosse (ISFA 2° année 2004-2005). L'office national interprofessionnel des vins –ONIVINS– est un établissement public à caractère industriel et commercial qui, entre autres, analyse l'évolution des marchés. Il se fait connaître par son site `http://www.onivins.fr` et publie un mensuel économique "ONIVINS-Infos". En collaboration avec l'INRA, il effectue régulièrement des enquêtes de consommation et rend publics les résultats.

Les personnes interrogées sont classées en trois catégories. Est **non consommateur**, une personne qui déclare ne jamais boire de vin. Est consommateur **régulier**, une personne qui déclare boire du vin chaque jour ou presque. Les autres sont des consommateurs **occasionnels**. Ces personnes interrogées sont d'autre part connues par leur sexe, âge et région. Dans le document

`http://www.onivins.fr/pdfs/218.pdf`

est proposé le pourcentage de buveurs réguliers par âge et date d'enquête, sous la forme :

	1980	1985	1990	1995	2000
<10	1.30	0.10	0.20	0.00	0.00
10-14	3.10	1.20	0.20	0.00	0.00
15-19	8.90	3.40	1.80	0.90	0.80
20-24	24.20	14.10	8.00	4.60	4.00
25-29	41.80	26.20	15.00	9.60	7.00
30-34	46.70	37.00	23.30	17.10	9.10
35-39	53.70	44.50	32.30	24.10	15.40
40-44	56.30	51.40	37.00	32.10	22.20
45-49	57.30	53.20	40.70	36.80	32.00
50-54	61.30	52.00	40.90	41.30	37.00
55-59	61.60	52.20	42.30	45.10	33.30
60-64	61.80	50.60	44.40	45.70	43.80
65-69	62.10	54.10	47.30	47.70	42.90
70-74	60.60	53.00	44.40	51.20	46.30
>75	60.10	51.30	48.30	49.90	42.70

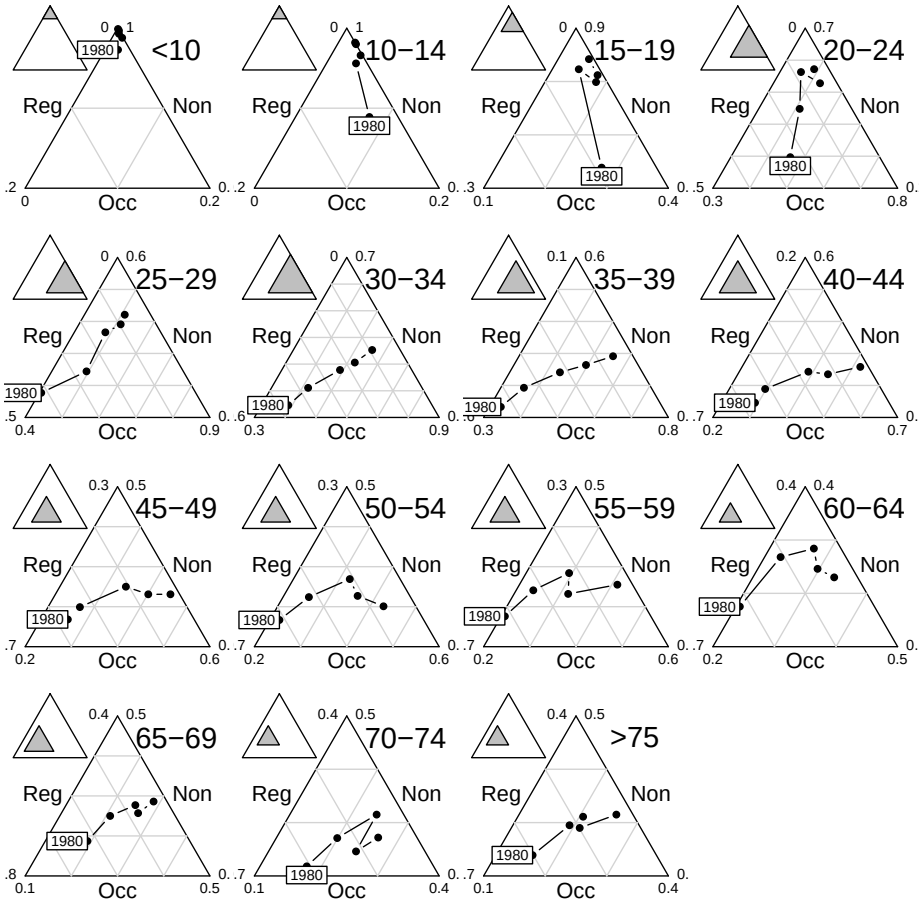
Les buveurs occasionnels des mêmes catégories sont représentés par :

	1980	1985	1990	1995	2000
<10	1.40	1.10	0.50	0.30	0.10
10-14	8.00	3.20	3.20	1.80	2.00
15-19	27.20	14.30	18.30	17.90	15.00
20-24	46.10	41.10	35.70	42.60	38.80
25-29	40.50	49.40	48.40	51.30	50.90
30-34	38.70	41.90	48.90	52.30	55.60
35-39	33.00	36.30	43.60	49.50	55.40
40-44	29.10	29.70	38.70	44.40	52.00
45-49	25.80	26.90	34.30	40.10	44.90
50-54	22.10	25.60	32.20	36.00	42.90
55-59	20.80	23.70	29.30	31.70	41.20
60-64	20.60	22.60	27.20	29.70	33.20
65-69	19.10	20.90	25.00	26.60	28.50
70-74	17.60	19.90	24.10	24.20	26.50
>75	16.00	19.20	20.60	21.10	25.80

On en déduira aisément la fréquence des non consommateurs. Les données sont disponibles par :

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/onivins.rda"))
```

Représenter la variabilité de la consommation de vin vue à travers ces enquêtes. Reconstituer laÒ figure suivante et étudier l'évolution de l'évolution ! Vous pensez à un autre point de vue ?



11 Problème : bière, vin ou alcool ?

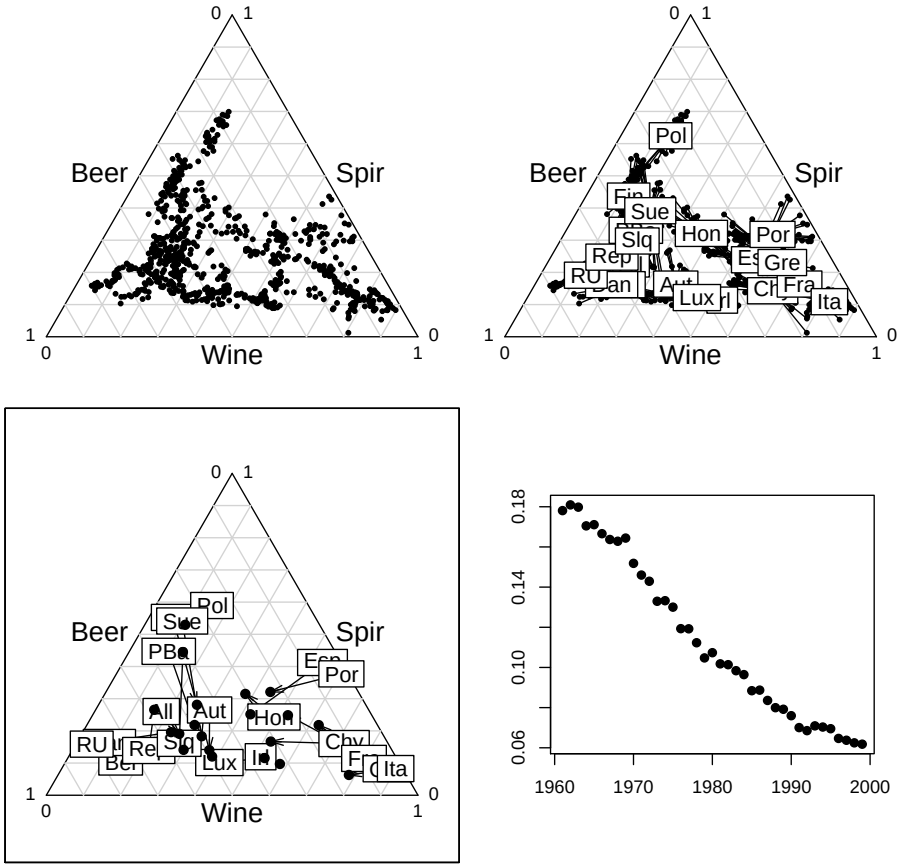
Pour achever cette introduction, sans sortir d'un contexte bien connu, utilisons le remarquable jeu de données sur la consommation d'alcool dans les différents pays d'Europe préparé par Christine Comptdaer et présenté dans :

`http://pbil.univ-lyon1.fr/R/pps/pps066.pdf`

Trois tableaux croisent 20 pays et 39 années pour la consommation individuelle de bière, vin et spiritueux. 3 produits, 20 pays, 39 années définissent un cube de données dans lequel on peut mettre en évidence une information considérable. On s'intéresse ici à une seule question. Chaque pays, chaque année présente un profil moyen de consommation : la quantité ne nous concerne pas, bien qu'il s'agisse d'un point de vue fort important. Chaque valeur est une consommation moyenne en litres d'alcool et on peut comparer les trois grands types de boissons qui contiennent cet alcool. Les sources et les détails sont disponibles dans la fiche citée. Pour disposer de cet exemple, utiliser :

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/pps066.rda"))
```

Reconstituer la figure suivante. On y voit la variabilité totale espace et temps confondu du profil de consommation. Cette variabilité forte induit la représentation par pays qui met en évidence les pays méditerranéens producteurs et consommateurs de vin et des comportements variés dans le reste de l'Europe. On cherche alors à montrer l'évolution de 1966 à 1999. Elle est également forte et ressemble à une homogénéisation. On calcule alors la variabilité de la position de l'ensemble des pays pour une année donnée. On appelle cette quantité la variance généralisée ou inertie. Elle diminue régulièrement. Naturellement, l'analyse aurait du aborder la question de la quantité d'alcool consommée, de sa variation entre pays et entre années, voire la variation de l'hétérogénéité de la consommation. On observe ici une évolution de structure et non de quantité. Passer de la variable à la structure est le propre de l'analyse géométrique des données.



Références

[1] B. Guinand, Y. Bouvet, and B. Brohon. Spatial aspects of genetic differentiation of the european chub in the rhone river basin. *Journal of Fish Biology*, 49 :714–726, 1996.

[2] P.M. Kroonenberg and R. Lombardo. Nonsymmetric correspondence analysis : a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, 34 :367–396, 1999.

[3] Ph. . Lebreton. *Les oiseaux nicheurs rhônalpins. Atlas ornithologique Rhône-Alpes*. Centre Ornithologique Rhône-Alpes, Université Lyon 1, 69621 Villeurbanne. Direction de la Protection de la Nature, Ministère de la Qualité de la Vie, 1977.